AD660661

No. K-54/67

Technical Memorandum

THE APPLICATION OF BLOCKING

IN REGRESSION ANALYSIS

Carl B. Bates

Computation and Analysis Laboratory

D D C

RECEIVED

NOV 13 1967

B

U. S. Naval Weapons Laboratory

Dahlgren, Virginia

Reproduced by the
CLEARINGHOUSE
for Federal Scientific & Technical
Information Springfield Va. 22151

37

U. S. NAVAL WEAPONS LABORATORY

TECHNICAL MEMORANDUM

August 1967

No. K-54/67

THE APPLICATION OF BLOCKING IN REGRESSION ANALYSIS

Carl B. Bates

Computation and Analysis Laboratory

Approved by:

RALPH A. NIEMANN
Director, Computation and
Analysis Laboratory

While the contents of this memorandum are considered to be correct, they are subject to modification upon further study.

Distribution of this document is unlimited.

TABLE OF CONTENTS

NWL Technical Memorandum No. K-54/67

* * * * * * * * * * * *

Initial Distribution:

Commander
Naval Air Systems Command (Code 604)
Washington, D. C. 20360 (2)

Commander
Naval Ordnance Systems Command (Code 9132)
Washington, D. C. 20360 (2)

Defense Documentation Center
Cameron Station
Alexandria, Virginia 22314 (20)

Initial Distribution (continued):

Local:

K	(1)
K-1	(1)
KE	(1)
KG	(1)
KP	(1)
KPO	(1)
KR	(1)
KRX	(1)
KRM	(20)
KRS	(1)
KRT	(1)
KRU	(1)
KW	(1)
KYD-1	(1)
T	(1)
TDM	(2)
TE	(1)
WDA	(1)
WHY	(1)
WWI	(1)
MAL	(6)
MAM	(1)

ABSTRACT

Occasionally the prediction equation obtained by conventional regression techniques is an unsatisfactory predictor because of its behavior over segments of the range of the independent variable(s). For such situations, a procedure is illustrated which has been found to yield a "better fit" than that obtained by conventional regression analysis. The procedure consists of segmenting the levels of the independent variable(s) into blocks and separately fitting each block. The separate fits, however, are obtained simultaneously and the end result is a single prediction equation. Numerical examples are given typifying regression analysis problems encountered in which the proposed procedure yields a "better fit". In each example, the proposed procedure of blocking in regression analysis is compared with conventional regression analysis. Extensions in the application of blocking in prediction problems and in comparative problems are briefly discussed.

I. INTRODUCTION

The principle of blocking in designed experiments conducted for comparative analysis purposes, namely the analysis of variance, is well established. However, the principle of blocking in experiments conducted for prediction purposes does not appear to be fully utilized. Indeed, a physical situation dictates the same restrictions upon experimentation conducted for prediction purposes as for comparative analysis purposes. That is, just as the analysis of variance is determined by the design of the experiment so should regression analysis be determined by the design of the experiment. In addition to the design of an experiment, another source of motivation for blocking in regression analysis is the demand from the experimenter for a "better fit". Often an experimenter's sole objective is to find a mathematical expression that "sufficiently fits" his data. That is, he is not interested in testing hypotheses concerning the parameters of some hypothesized model; instead, he is interested in the behavior of a mathematical function over a given range of the independent variable(s). This latter source of motivation initiated this report, and its objective is to illustrate the application of blocking by employing dummy variables in regression analysis to achieve a better fit than that obtained by conventional regression analysis.

The use of a dummy variable in regression analysis is not new. Many authors attach a dummy variable, which always takes the value of unity, to the constant (β_0) for notational convenience, especially when using matrix notation. Therefore, no pretension is made to the originality of using dummy variables in regression analysis; instead, an attempt is made

to extend the use of dummy variables in regression analysis. Likewise, the principle of blocking in regression analysis is not new; however, it has received surprisingly little attention in recent literature.

Suits (1957) uses dummy variables in regression analysis of independent variables which are partitioned (blocked) into mutually exclusive qualitative classifications. Klopfenstein (1964) stresses the utility of segmenting data in his discussion of the solution of the least squares approximation problem subject to a class of constraint conditions. Draper and Smith (1966) fit two linear trends to data which has been segmented into two blocks and illustrate the two cases of known and unknown point of intersection of the two trends. Smillie (1966) uses dummy variables to introduce qualitative variables into a regression function and gives a numerical example having a qualitative factor with two levels. The author of this report feels that a need exists for a more thorough exemplification of the utility of blocking in regression analysis than that illustrated in the current literature.

II. BACKGROUND

Knowledge of conventional regression analysis is assumed; therefore, neither the historical background nor the theory of regression analysis is discussed in detail. Instead, only definitions and/or explanations are given of the terminology and notation used later in the report.

In prediction problems concerning regression analysis involving a single dependent or response variable (y) and N independent variables

$(x_1, x_2, \cdots, x_N)$, the response variable is assumed to be normally distributed about the "true" response function (η) with common variance σ^2, where

$$\eta = \Phi(x_1, x_2, \cdots, x_N) \tag{1}$$

is linear in the parameters.

The objective is to determine a prediction equation which "fits" the given data with a prescribed degree of precision. This is accomplished by using a postulated model,

$$y = f(x_1, x_2, \cdots, x_N) + e, \tag{2}$$

to estimate η, where e is a random error. Assuming the general multiple linear regression model to be the postulated model, equation (2) is of the form

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_N x_N + e, \tag{3}$$

where

y = the dependent variable

x_ν = the ν^{th} independent variable; $\nu = 1, 2, \cdots, N$

β_0 = a constant

β_ν = the "true" partial regression coefficient of x_ν; $\nu = 1, 2, \cdots, N$

e = a random error.

Some of the independent variables may not be actually observed variables; for example, x_3 may equal x_1^2, x_4 may equal $x_1 x_2$, and so forth. In particular, in the case of a single independent variable (x), the postulated

model may be an N^{th} order polynomial and equation (3) becomes

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_N x^N + e. \tag{4}$$

Applying the least squares principle by minimizing the sum of squares of the deviations between the observed y_i values and the Y_i predictions yields unbiased estimates of the parameters of equation (3), where

$$Y_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + \cdots + b_N x_{Ni}; \; i = 1,2,\cdots,n, \tag{5}$$

and where n is the number of observed dependent variable values.

Concerning the distribution of the random errors (e_i), the usual assumptions of normally and independently distributed random errors with mean zero and variance σ^2 are assumed throughout the following discussion without further comment. For a complete discussion of the assumptions, see for example, Anderson and Bancroft (1952), Hald (1952), Bennett and Franklin (1954), or Johnson and Leone, Volume I (1964).

Criteria for judging the "goodness of fit" of a prediction equation are (or certainly should be) determined by the intended use of the prediction equation. Some of the more common criteria are based on the magnitude of the Coefficient of Multiple Determination (R^2), where R is the Multiple Correlation Coefficient; significance test of the "Lack of Fit"; and the magnitude of the residuals, $e_i = y_i - Y_i$ or $\bar{e}_i = \bar{y}_i - Y_i$. These criteria are used directly or indirectly in the **NUMERICAL EXAMPLES** Section where comparisons are made of blocking in regression analysis with conventional regression analysis.

III. NUMERICAL EXAMPLES

1. One Curvilinear Trend and One Linear Trend

Consider an experiment in which a single response was observed from each of 17 fixed levels of a given independent variable. The objective of the experiment was to determine a simple prediction equation (one containing as few terms as possible) for the response variable. For acceptance of a prediction equation, the residuals were to be within a prescribed tolerance, i.e., $|y_i - Y_i| \leq \delta$; $i = 1,2,\cdots,17$. The "true" response function was known to be monotonically increasing throughout the range of the independent variable. Additionally, the rate of change of the response function was increasing over a portion of the range of the independent variable, while the rate of change was nearly constant for the remainder of the range of the independent variable. The data was as follows.

Independent Variable (x), Dependent Variable (y)

x	y	x	y	x	y
1	0	7	37	13	51
2	1	8	40	14	51
3	6	9	43	15	55
4	10	10	43	16	56
5	18	11	46	17	59
6	28	12	47		

Least squares polynomials of increasing order were determined in the conventional manner. Prediction equations of the 8^{th} order and less failed to satisfy the specified tolerance. In addition, a prediction equation having more than five or six terms would have been impractical for the intended use of the prediction equation.

An examination of a plot of the data suggested that the transition from increasing to constant rate of change was between levels 6 and 8 of the independent variable. Therefore, the independent variable was segmented into two blocks, the first being from 1 through 7 and the second from 8 through 17. Constant, linear, and quadratic terms were fitted for the first block, and a linear term was fitted for the second block.

Before discussing the blocking procedure, the construction of the design matrix is briefly discussed. In the design matrix of TABLE I, x_1 and x_1^2 refer to the first block, x_2 refers to the second block, and x_3 represents the difference between the blocks. In the first block the elements of the x_1-column take the values of the original independent variable, and in the second block the elements of the x_1-column take the first value of the original independent variable in the second block. The elements of the x_1^2-column are the squares of the elements in the x_1-column. In the first block, the elements of the x_2-column take a zero; in the second block, they take the value of the original independent variable minus the first value of the original independent variable in the second block. The elements of the x_3-column are assigned a zero in the first block and assigned a one in the second block.

Note that the design matrix explained above and illustrated in TABLE I is not the only design matrix that could have been used. That is, the analyst is permitted flexibility in the construction of the design matrix. The elements of the columns referring to the blocks could have represented transformed or scaled values of the original independent variable. Similarly, the zero's and one's in the x_3-column could have been assigned differently. Naturally, a change in the construction of the design matrix changes the interpretation of the estimated regression coefficients.

TABLE I

DESIGN MATRIX AND RESPONSE DATA

Indep. Var. Index (i)	x		x_1	x_1^2	x_2	x_3	y
1	1		1	1	0	0	0
2	2		2	4	0	0	1
3	3	BLOCK I	3	9	0	0	6
4	4		4	16	0	0	10
5	5		5	25	0	0	18
6	6		6	36	0	0	28
7	7		7	49	0	0	37
8	8		8	64	0	1	40
9	9		8	64	1	1	43
10	10		8	64	2	1	43
11	11	BLOCK II	8	64	3	1	46
12	12		8	64	4	1	47
13	13		8	64	5	1	51
14	14		8	64	6	1	51
15	15		8	64	7	1	55
16	16		8	64	8	1	56
17	17		8	64	9	1	59

Four degrees of freedom were used for regression when blocking. Therefore, the 4^{th} order prediction equation, Y(C), obtained in the conventional manner is compared with the prediction equation, Y(B), obtained by blocking. The two prediction equations are:

$$Y(C) = - 0.2115 - 3.0124x + 2.3254x^2 - 0.2169x^3 + 0.0061x^4$$

$$Y(B) = - 0.5714 - 0.6310x_1 + 0.8690x_1^2 + 2.0667x_2 - 10.2000x_3$$

The MS(Lack of Fit) has been reduced by one-sixth as can be seen in the following ANOVA TABLE.

ANOVA TABLE

		CONVENTIONAL		BLOCKING	
Source	DF	SS	MS	SS	MS
Regression	4	6472.249	1618.062	6525.430	1631.358
Lack of Fit	12	62.810	5.234	9.629	0.802
Total	16	6535.059		6535.059	

Figure 1 shows a plot of the data, the 4^{th} order Y(C), and Y(B). The dashed portion of Y(B) between the two blocks illustrates the interpretation of the regression coefficient of x_3. The estimated regression coefficient (-10.2000) of x_3 is the vertical shift between the two trends of Y(B) at the first level of the second block. That is, Y(B) given $x_1=8$, $x_2=0$, $x_3=1$ minus Y(B) given $x_1=8$, $x_2=0$, $x_3=0$ equals -10.2000.

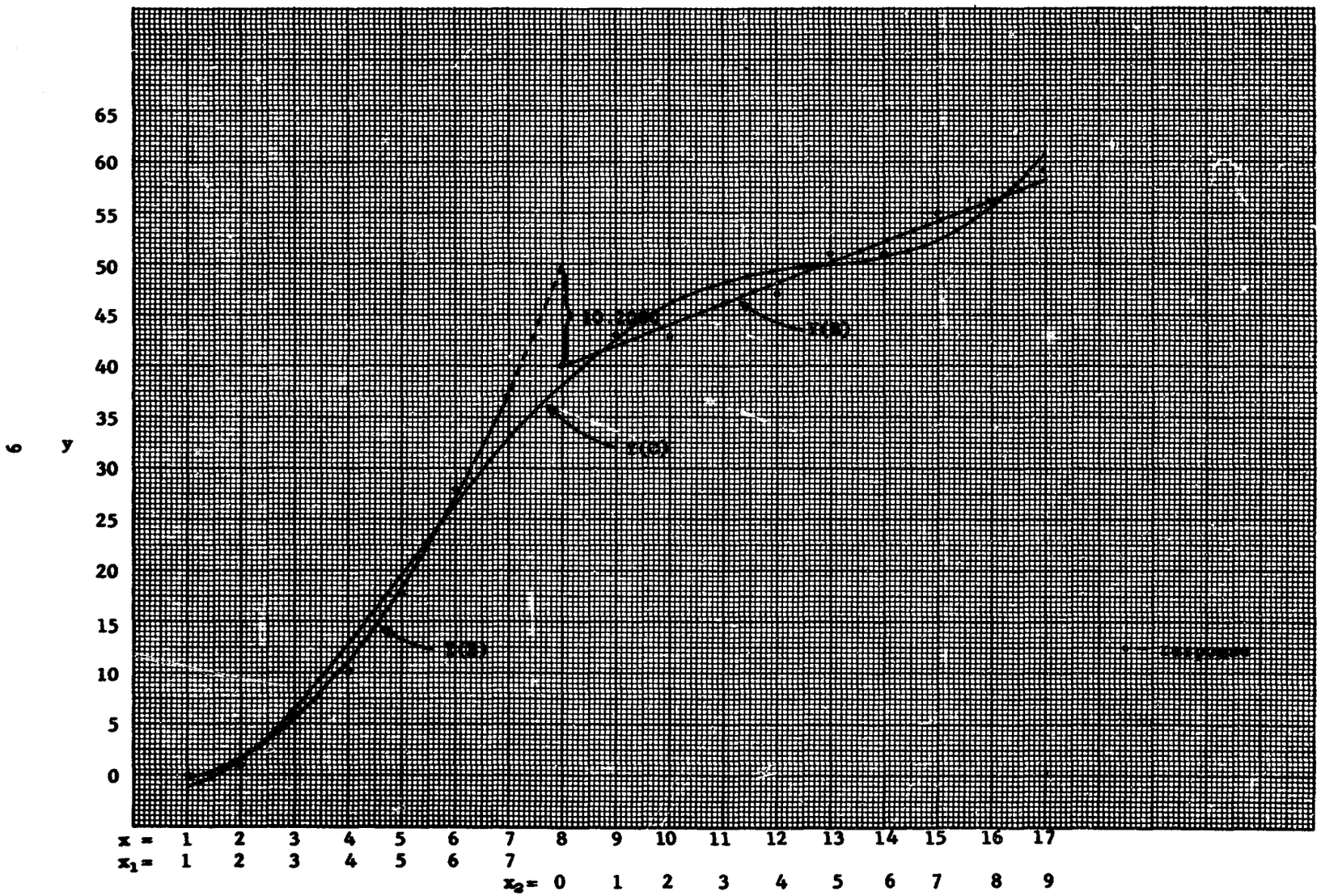
65
60
55
50
45
40
35
30
25
20
15
10
5
0
y
x(3)
x(2)
x(1)
x = 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17
x1 = 1 2 3 4 5 6 7
x2 = 0 1 2 3 4 5 6 7 8 9

Figure 1

A comparison of the residuals of Y(C) with the residuals of Y(B) shows that the $e_i(C)$ range from -2.86 to 4.21 while the $e_i(B)$ range from -1.20 to 1.13, where $e_i(C) = y_i - Y_i(C)$ and $e_i(B) = y_i - Y_i(B)$. Figure 2 shows a comparison of the residuals.

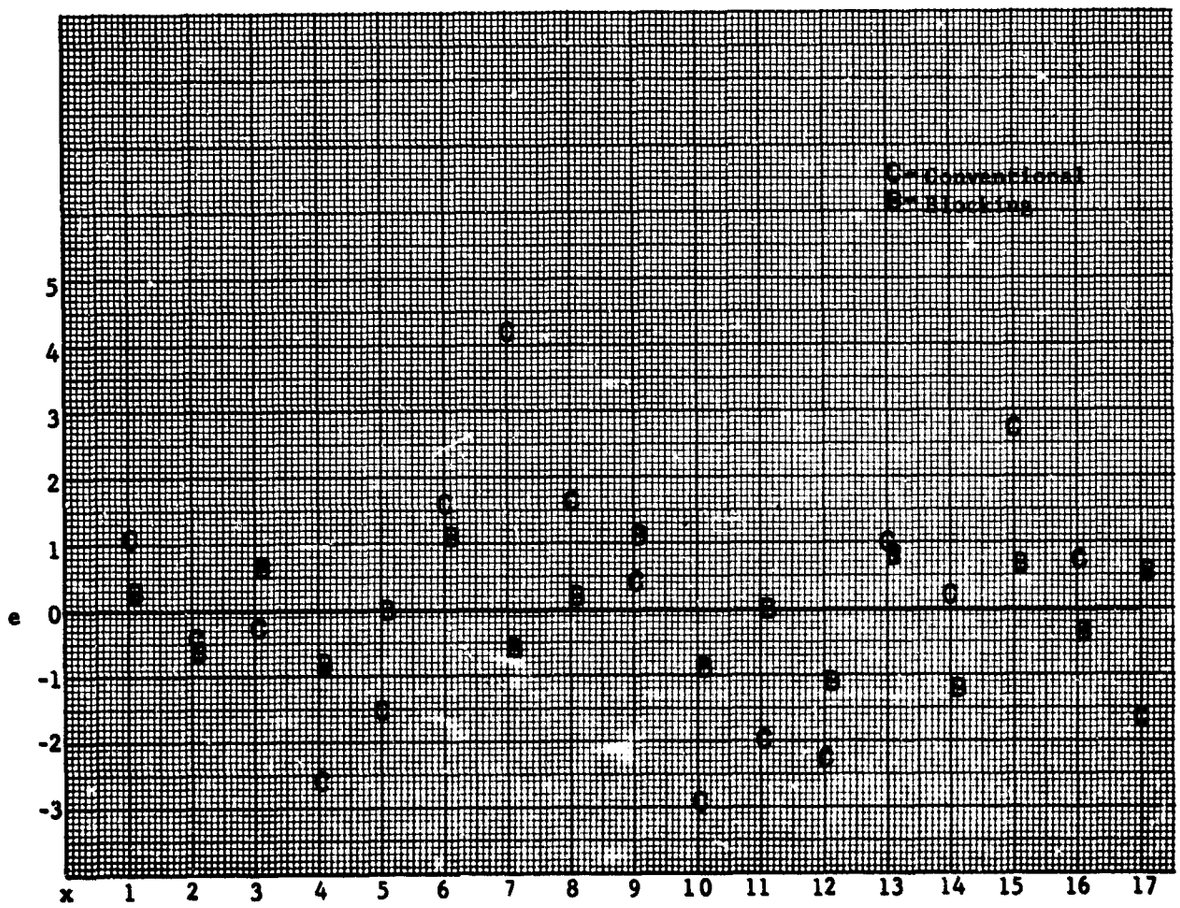

Figure 2

2. Two Curvilinear Trends

Consider another experiment in which 35 measured responses were obtained from 16 fixed levels of an independent variable. Again, the objective of the experiment was to obtain a simple prediction equation for the response variable. In addition, the prediction equation must possess certain characteristics, the most important being that it yield non-negative predictions for dependent variable values within the range of the experiment. Also, the true response function was known to be unimodal, and was known to be monotonically decreasing for increasing independent variable values to the right of the stationary point. The data was as follows.

Independent Variable (x)	Dependent Variable (y)		
1.0	0.5	1.0	1.5
1.5	6.0	8.0	
2.0	10.5	11.0	11.5
2.5	12.0	13.0	14.0
3.0	14.0	15.5	
3.5	15.0	16.0	
4.0	15.0	16.0	17.0
4.5	15.0	15.0	16.0
5.0	13.5	15.5	
5.5	10.5	11.0	11.0
6.0	6.0	8.0	
7.0	4.0	4.5	
8.0	2.5	3.0	
10.0	1.5		
15.0	0.7		
20.0	0.1		

Least squares fits were performed in the conventional manner, obtaining polynomial expressions in the independent variable. Prediction equations of the 9^{th} order and less were found to be unsatisfactory predictors. All curvilinear prediction equations yielded some negative predictions corresponding to x-values within the range of the experiment.

An examination of a plot of the data showed that in the range of 1 to 6 of the independent variable, the response trend was curvilinear and concave downward. But in the range of 6 to 20, the response trend was curvilinear, concave upward, and asymptotic to the x-axis as x increased. That is, the first trend appears as a portion of a parabola opening downward, while the second trend appears as a portion of a parabola opening to the right. Therefore, the independent variable was segmented into two blocks. For the first block linear and quadratic terms were included, and for the second block linear and square root terms were included. The design matrix is shown in TABLE II.

Because five degrees of freedom were used for regression when blocking, the 5^{th} order prediction equation, Y(C), obtained in the conventional manner is compared with Y(B) obtained by blocking:

$$Y(C) = -20.2869 + 26.9623x - 6.8518x^2 + 0.7012x^3 - 0.0319x^4 + 0.0005x^5$$

$$Y(B) = -10.8405 + 14.2910x_1 - 1.8808x_1^2 + 0.1773x_2 - 1.7921\sqrt{x_2} + 7.2348x_3$$

TABLE II

DESIGN MATRIX AND RESPONSE DATA

Indep. Var. Index (i)	x		x_1	x_1^2	x_2	$\sqrt{x_2}$	x_3	y		
1	1.0		1.0	1.00	0	0	0	0.5	1.0	1.5
2	1.5		1.5	2.25	0	0	0	6.0	8.0	
3	2.0		2.0	4.00	0	0	0	10.5	11.0	11.5
4	2.5		2.5	6.25	0	0	0	12.0	13.0	14.0
5	3.0	B L O C K	3.0	9.00	0	0	0	14.0	15.5	
6	3.5		3.5	12.25	0	0	0	15.0	16.0	
7	4.0		4.0	16.00	0	0	0	15.0	16.0	17.0
8	4.5	I	4.5	20.25	0	0	0	15.0	15.0	16.0
9	5.0		5.0	25.00	0	0	0	13.5	15.5	
10	5.5		5.5	30.25	0	0	0	10.5	11.0	11.0
11	6.0		6.0	36.00	0	0	0	6.0	8.0	
12	7.0		7.0	49.00	0	0	1	4.0	4.5	
13	8.0	B L O C K	7.0	49.00	1	1	1	2.5	3.0	
14	10.0		7.0	49.00	3	1.732	1	1.5		
15	15.0		7.0	49.00	8	2.828	1	0.7		
16	20.0	II	7.0	49.00	13	3.606	1	0.1		

A comparison of the "Lack of Fit" of Y(C) and Y(B) can be seen from the ANOVA TABLE below. The MS[Lack of Fit of Y(B)] is approximately one-fifth as large as the MS[Lack of Fit of Y(C)]. If a test were performed, the MS[Lack of Fit of Y(C)] would be found to be significant at the 0.01-level of significance, while the MS[Lack of Fit of Y(B)] is obviously not significant.

ANOVA TABLE

		CONVENTIONAL		BLOCKING	
Source	DF	SS	MS	SS	MS
Regression	5	1037.961	207.592	1065.681	213.136
Lack of Fit	10	34.407	3.441	6.687	0.669
Within	19	13.708	0.721	13.708	0.721
Total	34	1086.076		1086.076	

Figure 3 shows a plot of the data and the two prediction equations. Note that Y(C) yields negative values at x = 10,11,12,17,18,19. This, in addition to being a "poor fit" at x = 10, illustrates the danger of interpolation when the levels of the independent variable are unequally weighted and/or nonequidistant. Again, the estimated regression coefficient (7.2348) of x_3 is the vertical shift between the two trends of Y(B) at the first level of the second block. Figure 3 also contains a plot of the $\overline{e}_i = \overline{y}_i - Y_i$ differences, i.e., $\overline{e}_i(C) = \overline{y}_i - Y_i(C)$ and $\overline{e}_i(B) = \overline{y}_i - Y_i(B)$. These differences along with their corresponding predicted values are tabulated in TABLE III which shows the range of $\overline{e}_i(C)$ to be -2.08 to 1.89, while the range of $\overline{e}_i(B)$ is -0.64 to 0.91.

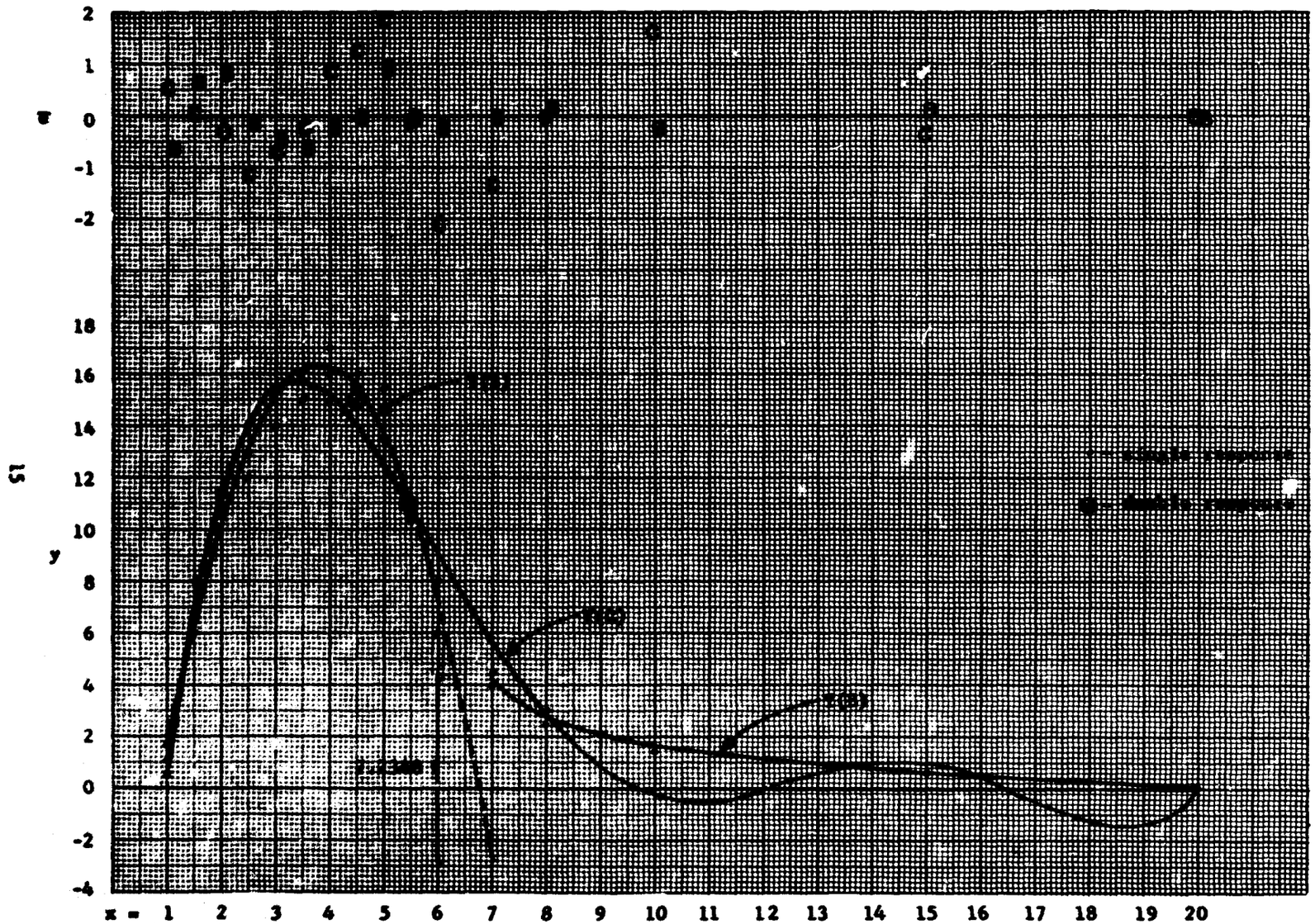
2
1
0
-1
-2
18
16
14
12
10
8
6
4
2
0
-2
-4
y
x = 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
x1 = 1 2 3 4 5 6
x2 = 0 1 2 3 4 5 6 7 8 9 10 11 12 13

Figure 3

TABLE III

PREDICTED VALUES AND RESIDUALS

i	x_i	$\bar{y}_i$	Y_i (C)	Y_i (B)	$\bar{e}_i$ (C)	$\bar{e}_i$ (B)
1	1.0	1.00	0.49	1.57	0.51	-0.57
2	1.5	7.00	6.95	6.36	0.05	0.64
3	2.0	11.00	11.35	10.22	-0.35	0.78
4	2.5	13.00	14.06	13.13	-1.06	-0.13
5	3.0	14.75	15.41	15.11	-0.66	-0.36
6	3.5	15.50	15.70	16.14	-0.20	-0.64
7	4.0	16.00	15.19	16.23	0.81	-0.23
8	4.5	15.33	14.09	15.38	1.24	-0.05
9	5.0	14.50	12.61	13.59	1.89	0.91
10	5.5	10.83	10.89	10.87	-0.06	-0.04
11	6.0	7.00	9.08	7.20	-2.08	-0.20
12	7.0	4.25	5.60	4.27	-1.35	-0.02
13	8.0	2.75	2.75	2.66	0.00	0.09
14	10.0	1.50	-0.16	1.70	1.66	-0.20
15	15.0	0.70	0.97	0.62	-0.27	0.08
16	20.0	0.10	0.07	0.12	0.03	-0.02

3. Three Linear Trends

This numerical example illustrates an extension of the blocking principle to three blocks. The hypothetical example is for demonstration of the procedure instead of comparison of blocking with conventional regression analysis. Therefore, results are presented, and the comparison is left to the reader. The data is as follows.

Independent Variable (x), Dependent Variable (y)

x	y	x	y	x	y
1	3.5	8	7.0	15	4.5
2	4.5	9	7.0	16	5.5
3	4.5	10	7.5	17	6.0
4	5.0	11	7.5	18	7.0
5	5.5	12	7.5	19	7.5
6	6.0	13	8.5	20	8.5
7	6.0	14	3.5	21	9.5

As shown in TABLE IV, the data is divided into three blocks having seven, six, and eight levels, respectively.

TABLE IV

DESIGN MATRIX AND RESPONSE DATA

Indep. Var. Index (i)	x		x_1	x_2	x_3	x_4	x_5	y
1	1		1	0	0	0	0	3.5
2	2		2	0	0	0	0	4.5
3	3	BLOCK I	3	0	0	0	0	4.5
4	4		4	0	0	0	0	5.0
5	5		5	0	0	0	0	5.5
6	6		6	0	0	0	0	6.0
7	7		7	0	0	0	0	6.0
8	8		8	0	0	1	0	7.0
9	9	BLOCK II	8	1	0	1	0	7.0
10	10		8	2	0	1	0	7.5
11	11		8	3	0	1	0	7.5
12	12		8	4	0	1	0	7.5
13	13		8	5	0	1	0	8.5
14	14		8	6	0	1	1	3.5
15	15		8	6	1	1	1	4.5
16	16	BLOCK III	8	6	2	1	1	5.5
17	17		8	6	3	1	1	6.0
18	18		8	6	4	1	1	7.0
19	19		8	6	5	1	1	7.5
20	20		8	6	6	1	1	8.5
21	21		8	6	7	1	1	9.5

The two prediction equations are:

$$Y(C) = 5.0589 - 1.5715x + 0.6116x^2 - 0.0697x^3 + 0.0031x^4 + 0.00005x^5$$

$$Y(B) = 3.3571 + 0.4107x_1 + 0.2571x_2 + 0.8214x_3 + 0.2143x_4 - 4.7750x_6$$

The amount of variation "explained" by each prediction equation is evidenced in the following ANOVA TABLE.

ANOVA TABLE

		CONVENTIONAL		BLOCKING	
Source	DF	SS	MS	SS	MS
Regression	5	40.124	8.025	55.005	11.001
Lack of Fit	15	15.662	1.044	0.781	0.052
Total	20	55.786		55.786	

Figure 4 shows a plot of the data and the two prediction equations.

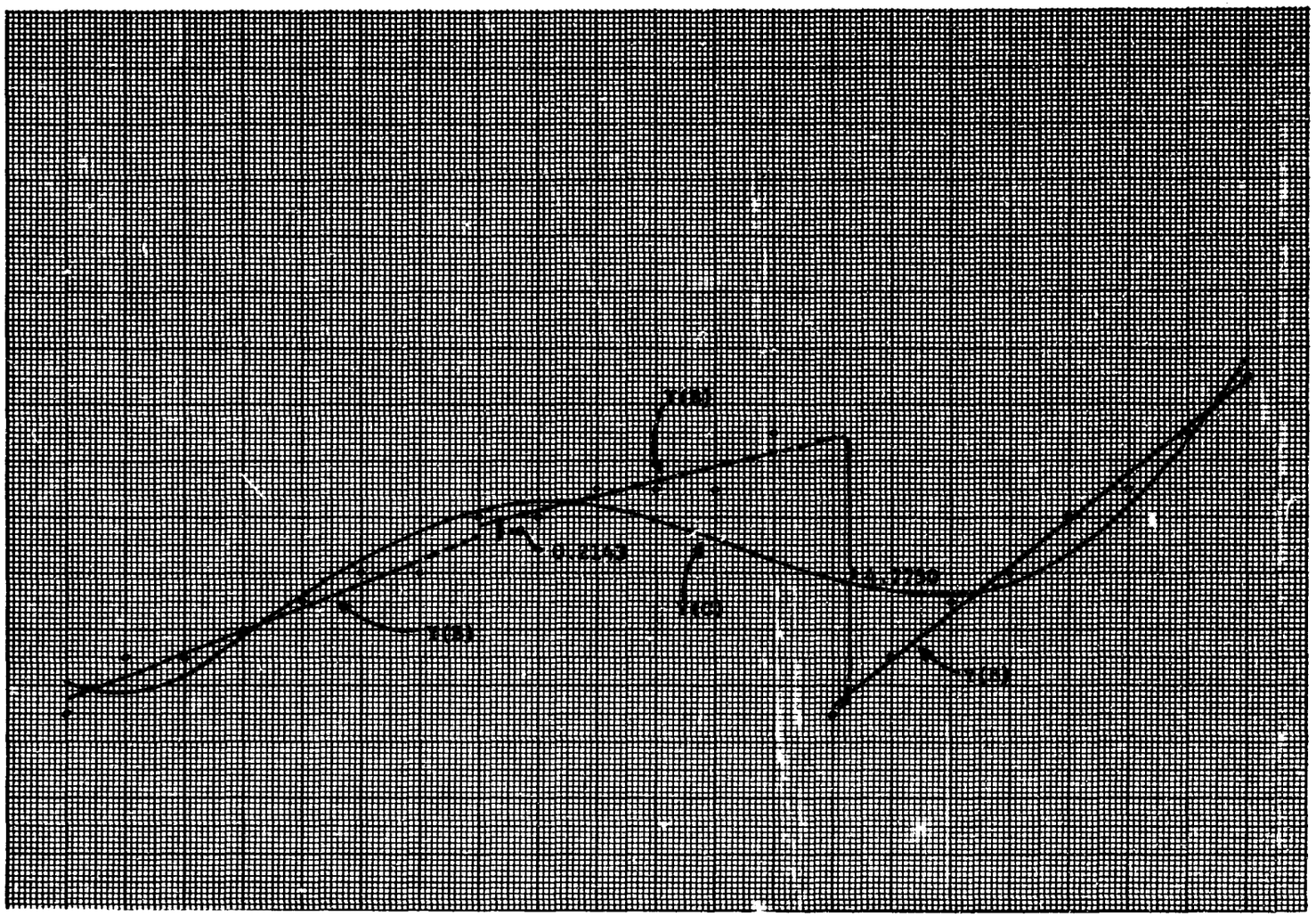
y
11
10
9
8
7
6
5
4
3
2
1
x = 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21
x1 = 1 2 3 4 5 6 7
x2 = 0 1 2 3 4 5
x3 = 0 1 2 3 4 5 6 7

Figure 4

IV. EXTENSION OF APPLICATION

1. Prediction Problems

The extension to K blocks is a straightforward generalization of the illustrations in Section III. An independent variable (x) having N levels may be segmented into K blocks as shown in TABLE V. The number of levels of the independent variable in the j^{th} block is N_j, where $\sum_{j=1}^{K} N_j = N$. Considering only linear terms in each block, the model is

$$y = \beta_0 + \sum_{j=1}^{K} \beta_j x_j + \sum_{j'=K+1}^{2K-1} \beta_{j'} x_{j'} + e. \qquad (6)$$

The estimates, b_j; $j = 1,2,\cdots,K$, of the parameters of equations (6) are the K slopes of the prediction equation; and the estimates, $b_{j'}$; $j' = K+1,K+2,\cdots,2K-1$, are the (K-1) vertical shifts between the K blocks. Naturally if desired, higher order terms of the type, $\beta_{j_{K_j}} x_j^{K_j}$; $j = 1,2,\cdots,K$; $K_j = 1,2,\cdots,N_j-1$, may be included in the model of equation (6). Further, the author sees no obvious complication in generalizing the above to multiple independent variables. The generalization appears to be an extension of the multiple regression approach to the analysis of variance illustrated by Brownlee (1960) or Draper and Smith (1966).

TABLE V

DESIGN MATRIX

		x_1	x_2	---	x_j	---	x_K	---	x_{K+j-1}	---	x_{2K-1}
B L O C K	I	1	0		0		0		0		0
		2	0		0		0		0		0
		3	0		0		0		0		0
		⋮	⋮	⋯	⋮	⋯	⋮	⋯	⋮	⋯	⋮
		x_{N_1-1}	0		0		0		0		0
		x_{N_1}									
B L O C K	II	x_{N_1+1}	0		0		0		0		0
		x_{N_1+1}	1		0		0		0		0
		x_{N_1+1}	2		0		0		0		0
		⋮	⋮	⋯	⋮	⋯	⋮	⋯	⋮	⋯	⋮
		x_{N_1+1}	x_{N_2-2}		0		0		0		0
		x_{N_1+1}	x_{N_2-1}		0		0		0		0
		⋮	⋮	⋯	⋮	⋯	⋮	⋯	⋮	⋯	⋮
B L O C K	J	x_{N_1+1}	x_{N_2}		0		0		1		0
		x_{N_1+1}	x_{N_2}		1		0		1		0
		x_{N_1+1}	x_{N_2}		2		0		1		0
		⋮	⋮	⋯	⋮	⋯	⋮	⋯	⋮	⋯	⋮
		x_{N_1+1}	x_{N_2}		x_{N_J-2}		0		1		0
		x_{N_1+1}	x_{N_2}		x_{N_J-1}		0		1		0
		⋮	⋮	⋯	⋮	⋯	⋮	⋯	⋮	⋯	⋮
B L O C K	K-1	x_{N_1+1}	x_{N_2}		x_{N_J}		0		1		0
		x_{N_1+1}	x_{N_2}		x_{N_J}		0		1		0
		x_{N_1+1}	x_{N_2}		x_{N_J}		0		1		0
		⋮	⋮	⋯	⋮	⋯	⋮	⋯	⋮	⋯	⋮
		x_{N_1+1}	x_{N_2}		x_{N_J}		0		1		0
		x_{N_1+1}	x_{N_2}		x_{N_J}		0		1		0
B L O C K	K	x_{N_1+1}	x_{N_2}		x_{N_J}		0		1		1
		x_{N_1+1}	x_{N_2}		x_{N_J}		1		1		1
		x_{N_1+1}	x_{N_2}		x_{N_J}		2		1		1
		⋮	⋮	⋯	⋮	⋯	⋮	⋯	⋮	⋯	⋮
		x_{N_1+1}	x_{N_2}		x_{N_J}		x_{N_K-2}		1		1
		x_{N_1+1}	x_{N_2}		x_{N_J}		x_{N_K-1}		1		1

2. Comparative Problems

In addition to the application of blocking in prediction problems, the procedure has application in the analysis of variance of both crossed and nested classifications. As an example of the application in crossed classifications, consider a simple 2X2X2 classification. The ANOVA model may be written as

$$y = \mu + a_\alpha + b_\beta + c_\gamma + ab_{\alpha\beta} + ac_{\alpha\gamma} + bc_{\beta\gamma} + abc_{\alpha\beta\gamma} + e. \qquad (7)$$

The corresponding REGRESSION model may be written as

$$y = \beta_0 + \sum_{\nu=1}^{7} \beta_\nu x_\nu + e. \qquad (8)$$

Applying regression analysis by using the design matrix of TABLE VI yields the analysis of variance for the three factor crossed classification.

TABLE VI

DESIGN MATRIX FOR A 2X2X2 CROSSED CLASSIFICATION

x_1	x_2	x_3	$x_4 = x_1x_2$	$x_5 = x_1x_3$	$x_6 = x_2x_3$	$x_7 = x_1x_2x_3$
1	1	1	1	1	1	1
1	1	2	1	2	2	2
1	2	1	2	1	2	2
1	2	2	2	2	4	4
2	1	1	2	2	1	2
2	1	2	2	4	2	4
2	2	1	4	2	2	4
2	2	2	4	4	4	8

Note: TABLE VI is an illustration of blocking applied to three independent variables (x_1 is segmented into two blocks, x_2 is segmented into two blocks within each block of x_1, and x_3 is segmented into two blocks within each block of x_2).

The correspondence of the analysis of variance for the models of equations (7) and (8) is illustrated in the following table.

ANOVA SOURCE COMPARISON

ANOVA MODEL SOURCE	REGRESSION MODEL SOURCE	DF
A	Due to $b_1 \mid b_0$	1
B	Due to $b_2 \mid b_0, b_1$	1
C	Due to $b_3 \mid b_0, b_1, b_2$	1
AB	Due to $b_4 \mid b_0, b_1, b_2, b_3$	1
AC	Due to $b_5 \mid b_0, b_1, b_2, b_3, b_4$	1
BC	Due to $b_6 \mid b_0, b_1, b_2, b_3, b_4, b_5$	1
ABC	Due to $b_7 \mid b_0, b_1, b_2, b_3, b_4, b_5, b_6$	1

For an illustration of the application of blocking in nested classifications, consider a two factor experiment in which a three level quantitative factor is nested within each of the three levels of a qualitative factor. The data is displayed in TABLE VII.

TABLE VII

DATA TABLE FOR A NESTED CLASSIFICATION

Factor A								
1			2			3		
			Factor B within A					
1	2	3	4	5	6	7	8	9
1	3	4	4	5	7	5	6	6
2	4	5	5	6	8	6	7	7

The ANOVA model may be written as

$$y = \mu + a_\alpha + b_{\beta(\alpha)} + e. \qquad (9)$$

Factor A has (A-1) = 2 degrees of freedom; factor B(A) has (B-1)A = 6 degrees of freedom. Applying the usual ANOVA computational procedures to the data in TABLE VII gives the following ANOVA TABLE.

ANOVA TABLE

Source	DF	SS	MS
A	2	32.444	16.222
B(A)	6	20.000	3.333
Within	9	4.500	0.500
Total	17	56.944	

Before applying the proposed blocking procedure, the regression model corresponding to the ANOVA model of equation (9) is briefly discussed. The terms within the regression model, and consequently the columns of the design matrix, are arranged differently from the arrangement used in the preceding sections of this paper. This rearrangement of terms within the regression model is merely for convenience so that the terms referring to factor A precede the terms referring to factor B within A (as they appear in the ANOVA model of equation (9)). That is, the set of (K-1) terms represented by the third term of equation (6) appears immediately after the constant β_0. Consequently, the REGRESSION model is written as

$$y = \beta_0 + \underbrace{\beta_1 x_1 + \beta_2 x_2}_{\text{Factor A}} + \underbrace{\beta_{31} x_3 + \beta_{32} x_3^2 + \beta_{41} x_4 + \beta_{42} x_4^2 + \beta_{51} x_5 + \beta_{52} x_5^2}_{\text{Factor B within A}} + e. \quad (10)$$

The design matrix corresponding to equation (10) is shown in TABLE VIII.

The ANOVA resulting from application of the proposed blocking procedure is given in the REGRESSION ANOVA TABLE. Testing the three parameters (β_{32}, β_{42}, β_{52}) of the quadratic terms in equation (10) as "Lack of Fit", we conclude that the departure from linearity is not significant. That is, a prediction equation containing only linear terms of the

TABLE VIII

DESIGN MATRIX FOR A THREE LEVEL NESTED CLASSIFICATION

		x_1	x_2	x_3	x_3^2	x_4	x_4^2	x_5	x_5^2	y	
BLOCK	I	0	0	1	1	0	0	0	0	1	2
		0	0	2	4	0	0	0	0	3	4
		0	0	3	9	0	0	0	0	4	5
BLOCK	II	1	0	4	16	0	0	0	0	4	5
		1	0	4	16	1	1	0	0	5	6
		1	0	4	16	2	4	0	0	7	8
BLOCK	III	1	1	4	16	3	9	0	0	5	6
		1	1	4	16	3	9	1	1	6	7
		1	1	4	16	3	9	2	4	6	7

quantitative factor "adequately fits" the data in TABLE VII. The resulting linear prediction equation is

$$Y = 0.1667 - 1.8333x_1 - 3.1667x_2 + 1.5000x_3 + 1.5000x_4 + 0.5000x_5 .$$

REGRESSION ANOVA TABLE

Source	DF	SS	MS
b_1&b_2	2	32.444	16.222
b_{31}	1	9.000	9.000
b_{32}	1	0.333	0.333
b_{41}	1	9.000	9.000
b_{42}	1	0.333	0.333
b_{51}	1	1.000	1.000
b_{52}	1	0.333	0.333
Within	9	4.500	0.500
Total	17	56.944	

Figure 5 shows a plot of the prediction equation and illustrates the interpretation of the estimated regression coefficients.

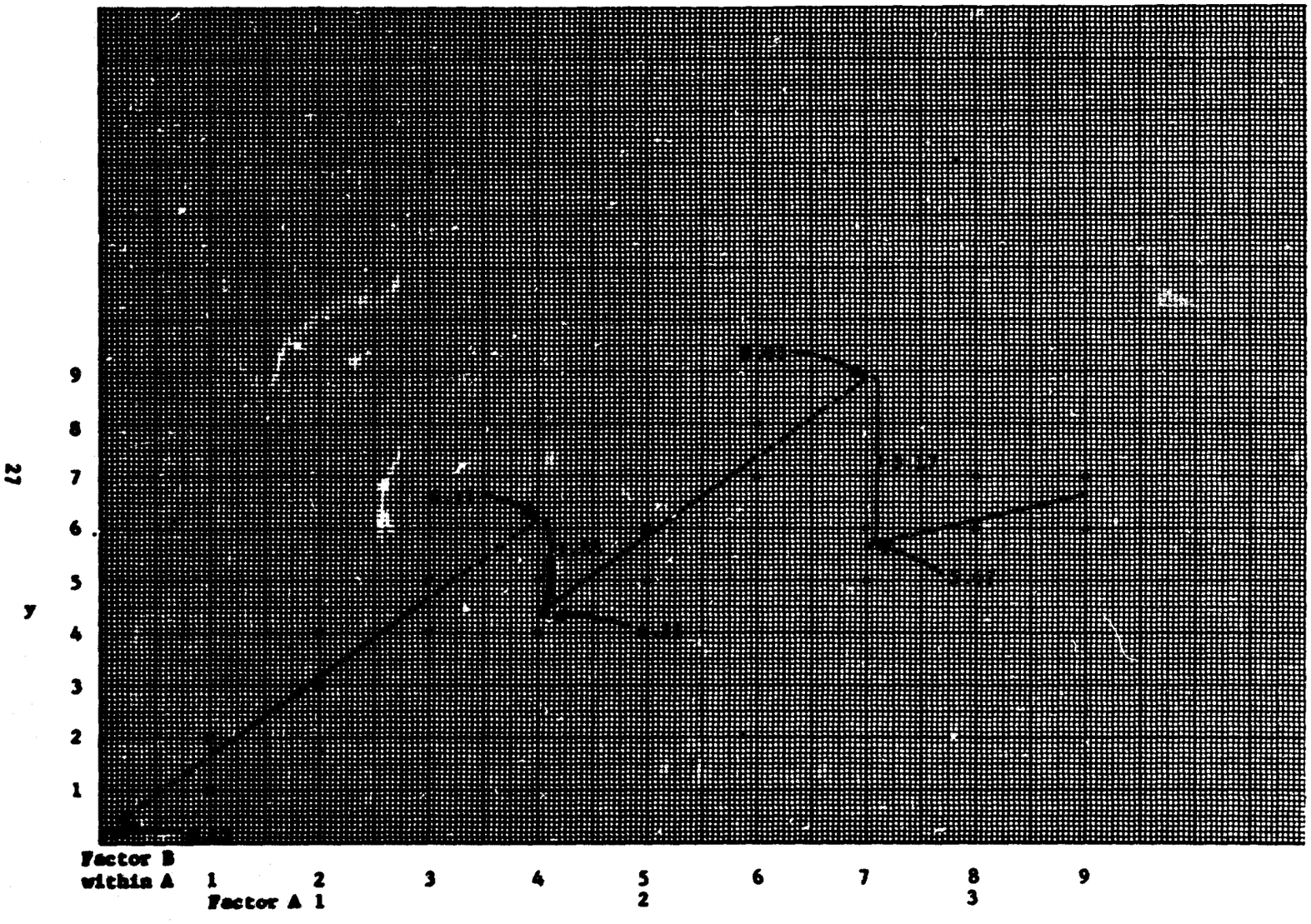

Figure 5

Note that application of the proposed blocking procedure enabled the simultaneous performance of an analysis of variance and a regression analysis. That is, in addition to the usual analysis of variance, a prediction equation was simultaneously determined.

In summary, the procedure of blocking in regression by using dummy variables provides the analyst much flexibility. This flexibility is due largely to the analyst's control of the construction of the design matrix. The elements of the design matrix may represent either original or transformed values of the original independent variable(s). Consequently, as illustrated in Section III.2, different transformations may be performed on different segments of the independent variable(s). In addition, the advantages afforded by employing orthogonal polynomials in regression analysis may be realized by constructing the columns of the design matrix to be orthogonal. Finally, with respect to the application to general analysis of variance problems, the author feels that the proposed procedure contained in this report could serve as a basis for a computer program applicable for the analysis of variance of both orthogonal and nonorthogonal designs having quantitative and/or qualitative factors in crossed and/or nested classifications.

V. REFERENCES

1. Anderson, R. L., and Bancroft, T. A. (1952), Statistical Theory in Research, McGraw-Hill Book Co., Inc., New York.

2. Bennett, C. A. and Franklin, N. L. (1954), Statistical Analysis in Chemistry and Chemical Industry, John Wiley and Sons, Inc., New York.

3. Brownlee, K. A. (1960), Statistical Theory and Methodology in Science and Engineering, John Wiley and Sons, Inc., New York.

4. Draper, N. R. and Smith, H. (1966), Applied Regression Analysis, John Wiley and Sons, Inc., New York.

5. Hald, A. (1952), Statistical Theory with Engineering Applications, John Wiley and Sons, Inc., New York.

6. Johnson, N. L. and Leone, F. C. (1964), Statistics and Experimental Design in Engineering and the Physical Sciences, Volumn I, John Wiley and Sons, Inc., New York.

7. Klopfenstein, R. W. (1964), Conditional Least Squares Polynomial Approximation, Mathematics of Computation, Vol. 18, pp. 659-662.

8. Smillie, K. W. (1966), An Introduction to Regression and Correlation, Academic Press Inc., New York.

9. Suits, D. B. (1957), Use of Dummy Variables in Regression Equations, Journal of the American Statistical Association, Vol. 52, pp. 548-551.

UNCLASSIFIED

Security Classification

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author)	2a. REPORT SECURITY CLASSIFICATION
- U. S. Naval Weapons Laboratory	UNCLASSIFIED
	2b. GROUP

3. REPORT TITLE

THE APPLICATION OF BLOCKING IN REGRESSION ANALYSIS

4. DESCRIPTIVE NOTES (Type of report and inclusive dates)

5. AUTHOR(S) (First name, middle initial, last name)

Carl B. Bates

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
August 1967	32	

8a. CONTRACT OR GRANT NO.	9a. ORIGINATOR'S REPORT NUMBER(S)
b. PROJECT NO.	TM K-54/67
c.	9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)
d.	

10. DISTRIBUTION STATEMENT

Distribution of this document is unlimited.

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY

13. ABSTRACT

Occasionally the prediction equation obtained by conventional regression techniques is an unsatisfactory predictor because of its behavior over segments of the range of the independent variable(s). For such situations, a procedure is illustrated which has been found to yield a "better fit" than that obtained by conventional regression analysis. The procedure consists of segmenting the levels of the independent variable(s) into blocks and separately fitting each block. The separate fits, however, are obtained simultaneously and the end result is a single prediction equation. Numerical examples are given typifying regression analysis problems encountered in which the proposed procedure yields a "better fit". In each example, the proposed procedure of blocking in regression analysis is compared with conventional regression analysis. Extensions in the application of blocking in prediction problems and in comparative problems are briefly discussed.

DD FORM 1 NOV 68 1473 (PAGE 1)

S/N 0101-807-6811

UNCLASSIFIED

Security Classification